Hierarchical Models: Random Effects

STAT 245

Random Effects

- What if residuals are *not independent*?
- In other words: we have "dependent data"...
- A solution (with many names): hierarchical models, multi-level models, random effects models, or mixed effects models.

Dataset

From: Falcone et al. 2017, *Diving behaviour of Cuvier’s beaked whales exposed to two types of military sonar*

Satellite tags were used to record dive data and movements of 16 Cuvier's beaked whales for up to 88 days each. The whales were incidentally exposed to different types of naval sonar exercises during the study. How did characteristics of their dives change during sonar exposure? We will look specifically at shallow dive duration as a response variable.

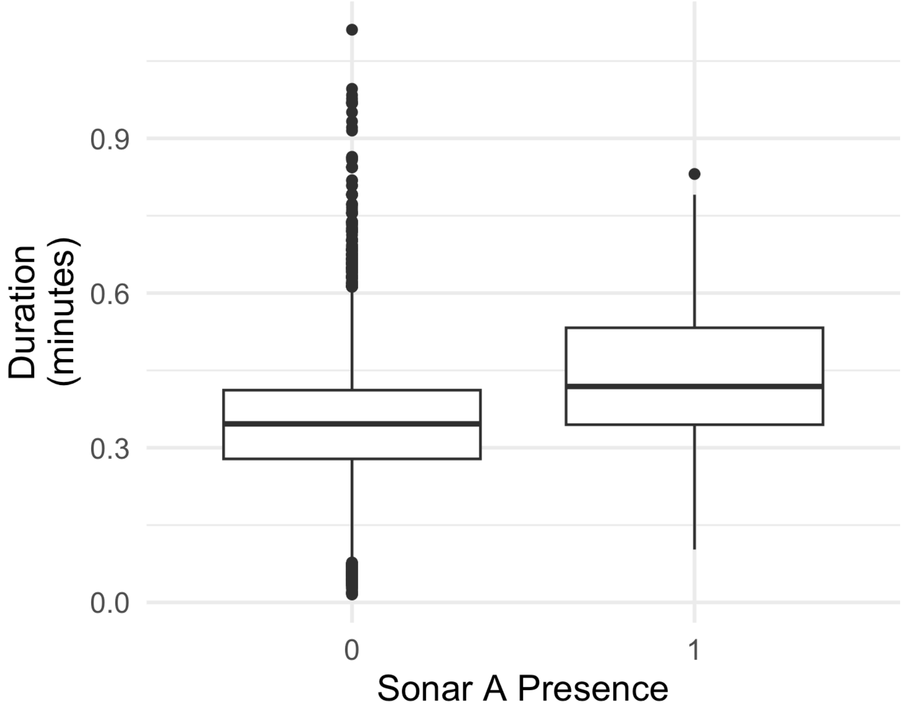
Dataset

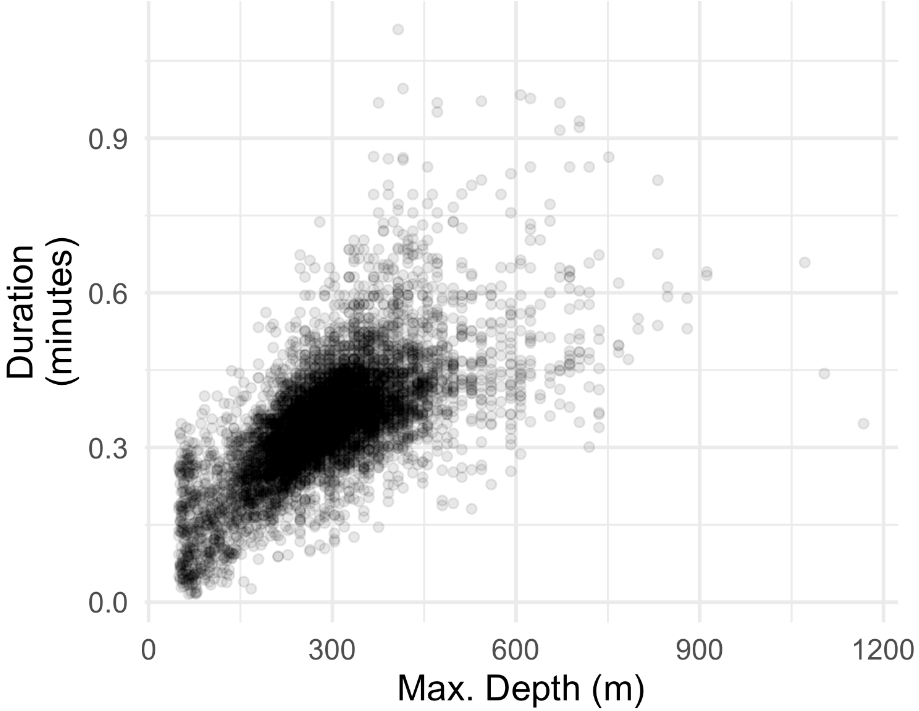
```r
zc_dives <- read.csv('http://sldr.netlify.com/data/zshal.csv') |>
  mutate(SonarA = factor(SonarA)) |>
  mutate(DurAvg = DurAvg / 60,
         DepthAvg = DepthAvg) |>
  rename(TimeOfDay = TransClass)
glimpse(zc_dives)
```

```
## Rows: 6,183
## Columns: 14
## $ TagID             <int> 14, 14, 14, 14, 14, 14, 14, 14, 14, 14, 14, 14, 14, …
## $ DurAvg            <dbl> 0.29300000, 0.32850000, 0.30183333, 0.34633333, 0.33…
## $ StartTime         <chr> "2011-01-06 20:45:30", "2011-01-06 22:13:23", "2011-…
## $ DepthAvg          <dbl> 335.5, 351.5, 287.5, 279.5, 359.5, 311.5, 263.5, 303…
## $ TimeOfDay         <chr> "Day", "Day", "Day", "Day", "Day", "Dusk", "Dusk", "…
## $ SonarA            <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ SonarB            <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ SonarAMinKm.fill  <dbl> 500, 500, 500, 500, 500, 500, 500, 500, 500, 500, 50…
## $ SonarBMinKm.fill  <dbl> 500, 500, 500, 500, 500, 500, 500, 500, 500, 500, 50
```

Plan/Exploration

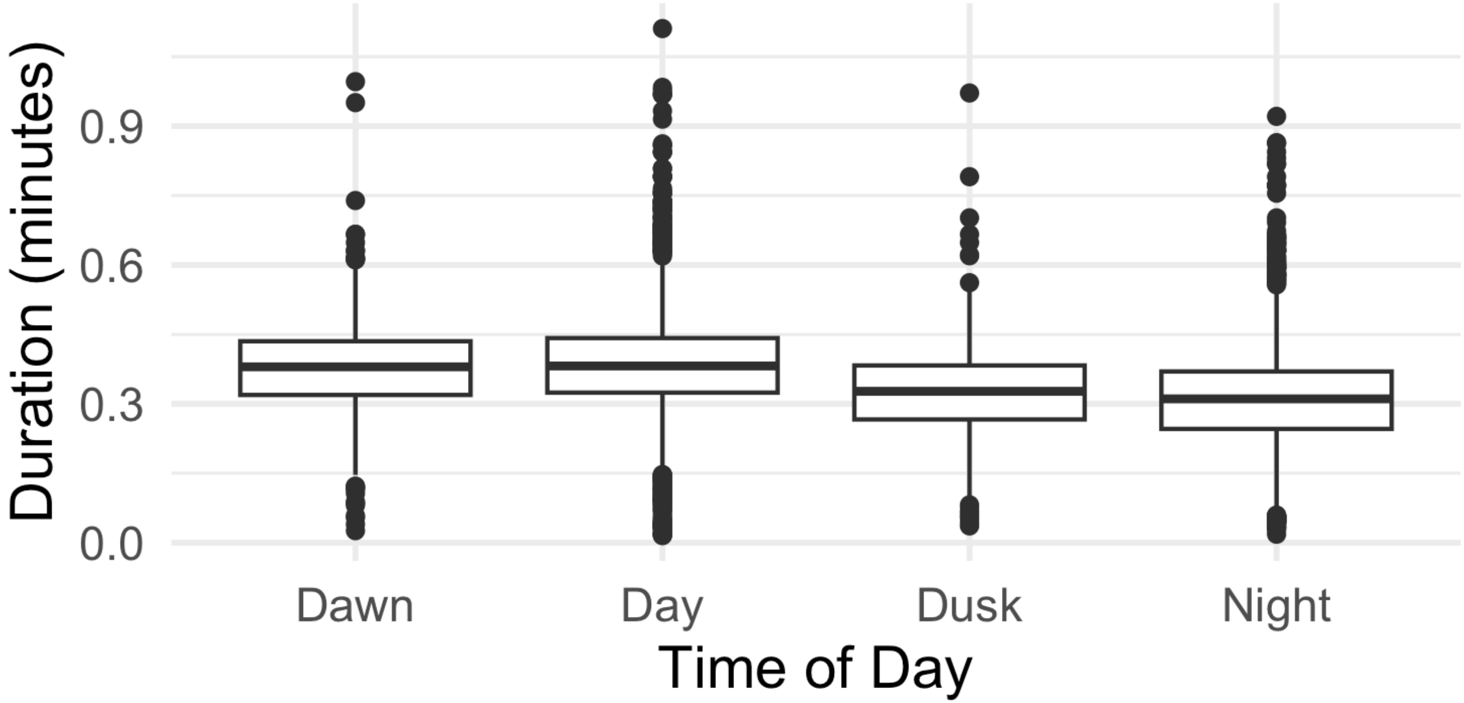
We are especially interested in how dive duration depends on sonar exposure. We also need to control for effects of other variables like depth and time of day.





Plan/Exploration

We are especially interested in how dive duration depends on sonar exposure. We also need to control for effects of other variables like depth and time of day.



What goes WRONG

if we use a `lm( )`?

(RE also work for count, binary data)

```
base_model <- lm(DurAvg ~ DepthAvg + TimeOfDay + SonarA,
                 data = zc_dives)
```

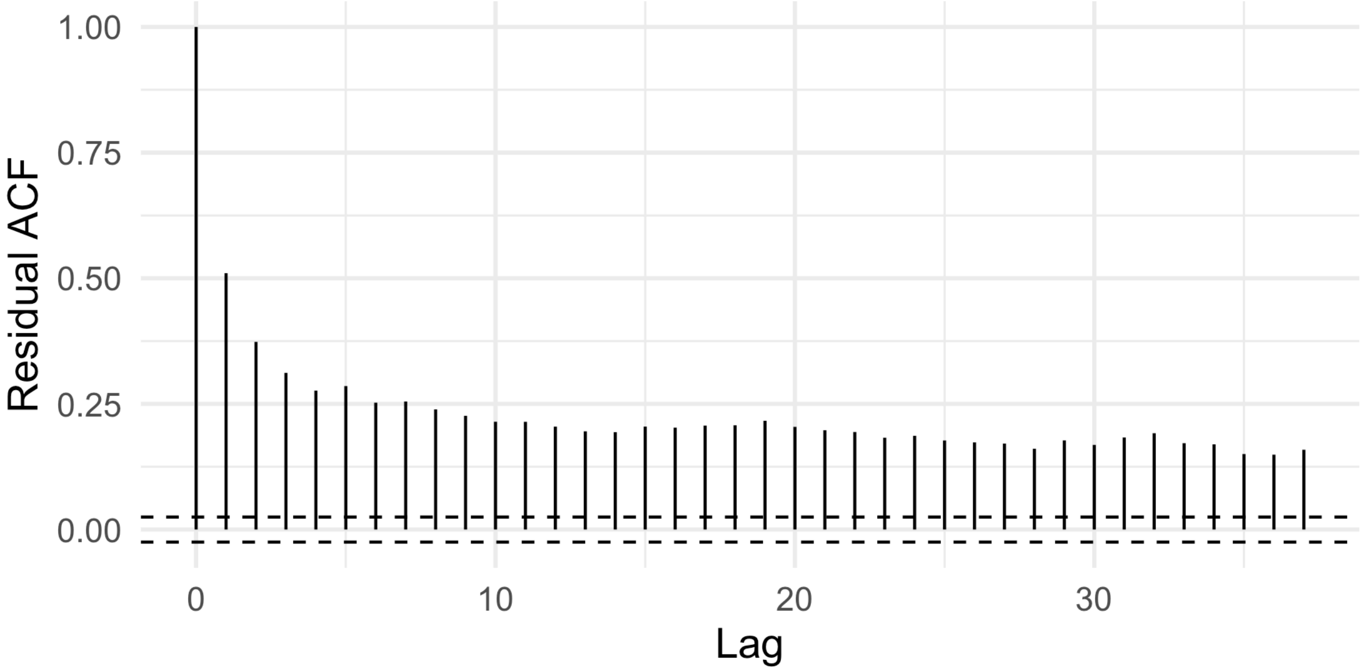
```r
mosaic::msummary(base_model)
```

```
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     1.833e-01  5.134e-03  35.712  < 2e-16 ***
## DepthAvg        6.446e-04  9.640e-06  66.863  < 2e-16 ***
## TimeOfDayDay   -1.073e-02  4.558e-03  -2.355   0.0186 *  
## TimeOfDayDusk  -3.356e-02  5.941e-03  -5.649 1.69e-08 ***
## TimeOfDayNight -4.046e-02  4.552e-03  -8.890  < 2e-16 ***
## SonarA1         6.273e-02  1.085e-02   5.781 7.80e-09 ***
## 
## Residual standard error: 0.08808 on 6177 degrees of freedom
## Multiple R-squared:  0.4757,	Adjusted R-squared:  0.4753 
## F-statistic:  1121 on 5 and 6177 DF,  p-value: < 2.2e-16
```

Model assessment

Data are *time series* so we are most suspicious about...

```
gf_acf(~base_model)
```



A Random Effects model

For multiple linear regression we would have:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_n x_n + \epsilon$$

Where $\epsilon \sim N(0, \sigma)$ are the normally distributed residuals with mean 0.

Now...

Random effect candidates?

- Thermal preference
- NY car crashes
- Sports votes
- PhD innovation
- Wood frog abnormality

The Formula

and the function: `glmmTMB()`.

Many also use `lme4::(g)lmer()`

- We add random effects to the model formula with:

$$+(1|\text{ variable})$$

Formula

Nested Random Effects

$$\text{response} \sim \ldots + (1|\ \text{variable1}\ /\ \text{variable2})$$

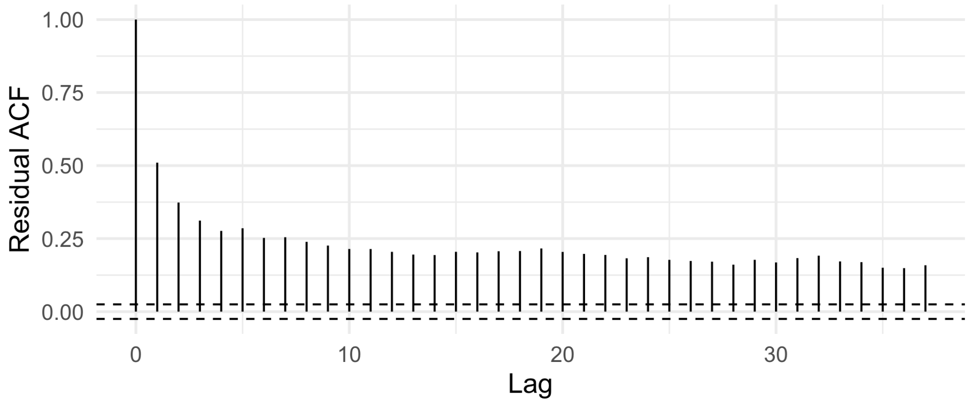
Random effect of individual whale

```
rem1 <- glmmTMB(DurAvg ~ DepthAvg + TimeOfDay +
                  SonarA + (1|TagID),
                data = zc_dives)
```

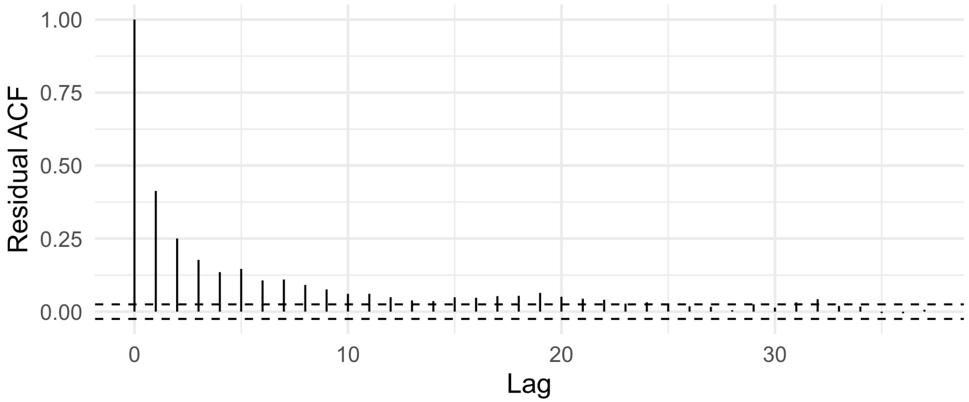
Model Assessment

Focus on ACF now as it was the problem for the lm()

```
gf_acf(~base_model)
```



```
gf_acf(~resid(rem1))
```



What else would we need to check?

What can we try next?

```
glimpse(zc_dives)
```

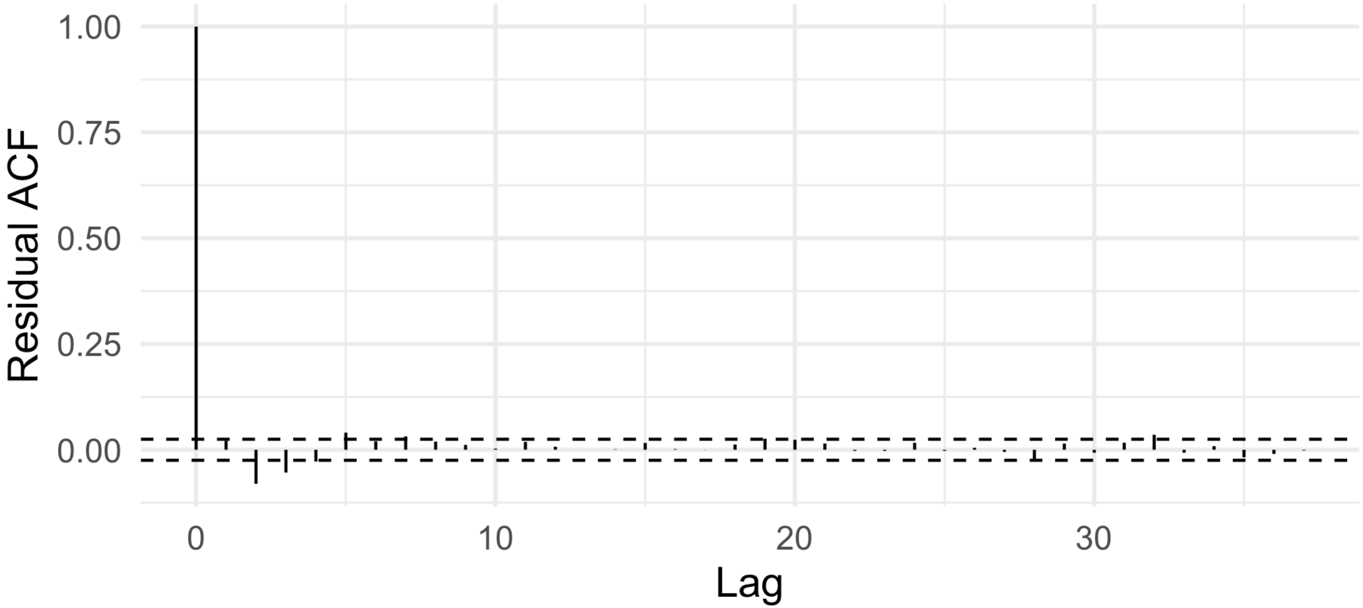
```
## Rows: 6,183
## Columns: 14
## $ TagID             <int> 14, 14, 14, 14, 14, 14, 14, 14, 14, 14, 14, 14, 14, …
## $ DurAvg            <dbl> 0.29300000, 0.32850000, 0.30183333, 0.34633333, 0.33…
## $ StartTime         <chr> "2011-01-06 20:45:30", "2011-01-06 22:13:23", "2011-…
## $ DepthAvg          <dbl> 335.5, 351.5, 287.5, 279.5, 359.5, 311.5, 263.5, 303…
## $ TimeOfDay         <chr> "Day", "Day", "Day", "Day", "Day", "Dusk", "Dusk", "…
## $ SonarA            <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ SonarB            <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ SonarAMinKm.fill  <dbl> 500, 500, 500, 500, 500, 500, 500, 500, 500, 500, 50…
## $ SonarBMinKm.fill  <dbl> 500, 500, 500, 500, 500, 500, 500, 500, 500, 500, 50…
## $ SonarAPercOL.fill <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ SonarBPercOL.fill <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ TagDay            <chr> "2011-01-06", "2011-01-06", "2011-01-06", "2011-01-0…
## $ Period            <chr> "(18,20]", "(20,22]", "(20,22]", "(20,22]", "[0,2]",…
## $ TagDayPeriod      <chr> "2011-01-06.(18,20]", "2011-01-06.(20,22]", "2011-01…
```

What can we try next?

```
rem2 <- glmmTMB(DurAvg ~ DepthAvg + TimeOfDay +
                  SonarA +
                  (1|TagID/TagDayPeriod),
                data = zc_dives)
```

Better?

```
gf_acf(~resid(rem2))
```



Comparison with `lm()`?

How does this model compare to the original linear regression model?

- Coefficient estimates?
- SEs?
- Additional stuff in the summary output?

Coefficients

Variable	lm() est.	RE est.	lm SE	RE SE
(Intercept)	0.183	0.177	0.005	0.009
DepthAvg	0.001	0.001	0.000	0.000
TimeOfDayDay	-0.011	-0.003	0.005	0.004
TimeOfDayDusk	-0.034	-0.021	0.006	0.006
TimeOfDayNight	-0.040	-0.038	0.005	0.004
SonarA1	0.063	0.048	0.011	0.012

Model Selection for Mixed Models

Standard likelihood-based model selection criteria? Well...yes, and no.

REML or ML?

Two different ways to fit these models:

ML

REML

*`glmmTMB()` **default is: REML = FALSE***

Selection

(one way)

```
zc_dives_noNA <- zc_dives |>
  drop_na(DepthAvg, TimeOfDay, SonarA, TagID, TagDayPeriod)
rem_sonar <- glmmTMB(DurAvg ~ DepthAvg + TimeOfDay + SonarA +
                       (1|TagID),
                     data = zc_dives_noNA)
rem_no_sonar <- glmmTMB(DurAvg ~ DepthAvg + TimeOfDay +
                       (1|TagID),
                     data = zc_dives_noNA)
BIC(rem_sonar, rem_no_sonar)
```

```
##              df       BIC
## rem_sonar     8 -13482.98
## rem_no_sonar  7 -13441.11
```

Random Slopes?

- What we just did ("random effects") also called "random intercept" model
- We allowed for an offset between the overall average predicted response value and that of an individual
- We did not allow the *slope* of the relationship with any of the predictor variables to vary randomly with individual.
- It is possible to do this, **But it often makes interpretation difficult**.

Sketch: random intercept vs. slope

Random slope = interaction between RE variable & predictor

Animation: http://mfviz.com/hierarchical-models/

Before random slopes...

Maybe Don't. Ask yourself:

- Do you really think that there is random variation in the *slope* of the predictor with the response?
- Is there a strong, clear overall effect and small variations in its magnitude between individuals?
- Will the relationship with a certain predictor have very strong and very different slopes for different individuals?
- Is dataset big enough?

Random Slopes Formula(s)

$$... + (\text{PredictorVariable} \mid \text{REGroupingVariable})$$

or equivalently

$$... + (1 + \text{PredictorVariable} \mid \text{REGroupingVariable})$$

or random slope *without* the corresponding random intercept **don't ever do it...**:

$$... + (0 + \text{PredictorVariable} \mid \text{REGroupingVariable})$$

Coming next: Prediction Plots

- Do we want to make predictions for "the average random effect grouping" (here, the average whale in the average time-block)?
- Or do we want to make predictions *averaged across* a whole population of random effect groups (all the whales at all times)?
- For a *linear* model, *the two are numerically equal* but for models with a link function *they are different and we have to choose.*